methods, data, analyses | Vol. 10(1), 2016, pp. 47-72 DOI: 10.12758/mda.2016.003

Privacy Concerns in Responses to Sensitive Questions. A Survey Experiment on the Influence of Numeric Codes on Unit Nonresponse, Item Nonresponse, and Misreporting

Felix Bader[1, 2], *Johannes Bauer*[1, 3], *Martina Kroher*[4] & *Patrick Riordan*[1]

1 *Department of Sociology, Ludwig-Maximilians-Universität München*
2 *Mannheim Centre for European Social Research (MZES)*
3 *Institute for Employment Research (IAB), Nuremberg*
4 *Institute for Sociology, Leibniz Universität Hannover*

Abstract

Paper-and-pencil surveys are a widely used method for gaining data. Numeric codes printed on the questionnaire are often a prerequisite for the use of scan software, which, in turn, permits a fast and efficient entering of the data from such surveys. However, printed numbers used for optical mark recognition on a questionnaire can provoke concerns about anonymity that may lead to unit nonresponse, item nonresponse, and misreporting.

To test this, we conducted an experiment in a mail survey on group-focused enmity, printing a scanner code on half of the questionnaires. Our results show no significant deviation concerning unit nonresponse. We find a higher item nonresponse and misreporting bias towards socially desirable answers in sensitive questions if the questionnaire is marked with a code. The influence of biased responses on regression results is minor. If the numeric code is brought to the respondents' attention in the cover letter, regression coefficients might be affected. Therefore we conclude that researchers should trade off these small biases against the usefulness of the code. From a methodological perspective, we recommend not to make a statement concerning the numeric code in the cover letter.

Our results are of relevance for researchers conducting paper-and-pencil surveys as well as for those analyzing data sets from these surveys. While this article analyzes biases caused by scanner codes, the results are potentially transferable to printed identification numbers used in panel studies, in survey experiments, or to match paradata or context data.

Keywords: questionnaire design, scanner codes, sensitive questions, tailored design, unit nonresponse, item nonresponse, misreporting

1 Introduction

In this study, we analyze the effects of numeric codes printed on paper questionnaires and their influences on respondent answers. These codes are often used in scan software for a fast and efficient entering of data. While we are primarily concerned with paper-and-pencil mail and interviewer surveys, with mixed-mode surveys becoming more prevalent the following results are also relevant for other survey types. This study focuses on scanner codes, but there are other potential purposes of such codes. They might be useful to identify respondents in panel studies or recognize treatment groups in survey experiments. Another application is the adding of paradata or context data. Regional identification numbers printed on the questionnaire, for example, can help to evaluate regional nonresponse and append geodata and other context data from external sources.

In paper-and-pencil surveys the respondent either completes the questionnaire herself or with an interviewer and the answers are filled in manually on a paper questionnaire. Subsequently this data needs to be digitalized in some way in order to enable researchers to efficiently analyze it. The digitalization can be done by manual input or by utilizing scan software.

The obvious advantages of scanning questionnaires is the considerable amount of time saved compared to the arduous procedure of manual data input and the higher quality of the produced data set. Also, when different people collaborate in the manual data processing, structural errors like different coding of missings and filter questions can become a problem. Consequently, scanning is advantageous in many respects and part of multiple recent and past surveys.

However, a problem when capturing data optically may arise due to the use of numeric codes, barcodes or QR codes, which are printed on the questionnaire. Prevalent software like Readsoft, TeleForm, EvaSys, and at least 20 other popular tools utilize optical mark recognition, which uses a printed code to identify a stored master form when processing the data. Nine member institutes of the German ADM-Sampling-System for Face-to-Face Surveys regularly apply paper-and-pencil surveys. An informal survey revealed that roughly half of these institutes use and recommend printed scanner codes.

Direct correspondence to
Johannes Bauer, Department of Sociology, LMU Munich, Konradstr. 6, 80801 Munich
E-Mail: johannes.bauer@lmu.de

Acknowledgements: We are grateful to the participants of the course in research practice "Methods of quantitative empirical social research 1" in 2013 who helped preparing the survey documents, generated the addresses, and delivered the questionnaires. Many thanks to Christian Ganser who provided valuable advice and support during all phases of the project. We also thank two anonymous reviewers for their helpful comments.

While scanner codes do not permit an identification of respondents, survey participants might feel that their anonymity is jeopardized by such numbers and react to a perceived breach of their anonymity in one of three ways: First, they might not participate in the survey at all (unit nonresponse). Second, they might participate, but decline answering sensitive questions (item nonresponse). Third, they might take part and answer even sensitive questions, but their answers might be biased by social desirability (misreporting). Misreporting is a general problem in surveys, especially in surveys with sensitive topics (Preisendörfer & Wolter, 2014; Wolter, 2012).

In this paper we present a survey experiment conducted to examine whether such codes lead to any of the three mentioned reactions by respondents. We delivered paper questionnaires on political opinions and group-focused enmity (GFE) to a random sample of the resident population of Munich. Respondents randomly received one of three versions: a) questionnaire without numeric code (these were carefully input by hand), b) questionnaire with numeric code but without specific statement concerning this code in the cover letter, and c) questionnaire with numeric code and a specific statement in the cover letter truthfully explaining the purpose of the code and that it cannot be used to jeopardize anonymity. The latter procedure is proposed by Dillman (2007) if numeric codes are inevitable. We compare unit nonresponse, item nonresponse, and answers to sensitive questions of these three groups as well as differences in results of typical GFE-regressions to determine the influence of such codes.

In the following section we outline the theoretical arguments relevant to our research question and discuss previous related research (section 2). In section 3 we present our data and the experimental design, before reporting central findings in section 4. We discuss these findings and their applicability to other topics and conclude the paper with some implications for practical research in section 5.

2 Theoretical and Empirical State of Research

There is an extensive theoretical and empirical literature on the factors influencing the response of survey participants. In this section we briefly outline the basic theoretical argument underlying our hypotheses and discuss empirical work directly related to the effects of survey design, sensitive questions and respondents' concern for the anonymity of their answers. For brevity's sake, this outline is limited to self-administered mail surveys, since this is the mode in question for this study. Regarding answers to sensitive questions, self-administered surveys tend to lead to less biased answers than interview-based surveys (Bradburn, Sudman, & Wansink, 2004; Richman, Kiesler, Weisband, & Drasgow, 1999; Stocké, 2004).

2.1 Tailored Design

The literature on the "tailored design method" (Dillman, 2007) has shown that in order to maximize response in mail surveys it is necessary to pay attention to every detail of the questionnaire, the cover letter, and all other elements submitted to the respondents (Babbie, 2013; de Leeuw & Hox, 2008; de Leeuw, Hox, & Huisman, 2003; Dillman, 1991, 2007, 2008). Although the efficiency of these elements has not been empirically settled (e.g. de Rada, 2005; Edwards et al., 2002) it is a practical default to assume that potential responders will react to various aspects of the survey materials.[1]

In the widely used model of rational respondents, survey participation and truthful answers hinge on the respondents' sense of benefits outweighing costs (Dillman, 1991, 2007; Tourangeau, Rips, & Rasinski, 2000). Warwick and Lininger have outlined factors on the one hand contributing to survey response and on the other hand preventing respondents from giving information as early as 1975 (see also Lessler & Kalsbeek, 1992). They describe how individuals are willing to share their experiences with interested listeners, as long as the questions are not sensitive and participation is not too costly in any other way.

Respondents can have major concerns about the anonymity of their data. This is why virtually all survey researchers make a statement on anonymity or confidentiality at some point, usually in the cover letter. Responders will feel even more anonymous, when there is no possible way to breech their anonymity: "If you make it a white envelope without any marks on it other than the address the questionnaire has to be sent to, and if the questionnaire does not contain any visible form of numbering, the respondent will feel freer to respond honestly to the questions" (Lensvelt-Mulders, 2008, p. 470). As described above, automated optical data capture using scanners makes some form of coding on questionnaires necessary. We aim at exploring whether these codes influence the responses.

2.2 Sensitive Questions, Privacy, and Nonresponse

Whether respondents give truthful answers to sensitive questions or not, is a classic issue in survey methodology (Barton, 1958; Benson, 1941; Hyman, 1944) and numerous more recent studies have shown that respondents tend to underreport socially undesired behavior and to overreport socially desired behavior (Barnett, 1998; Beyer & Krumpal, 2010; Kreuter, Presser, & Tourangeau, 2008; Lee, 1993; Tourangeau et al., 2000). Regarding the domain of misreporting in surveys, research on sensitive questions is a very important matter (for reviews see Krumpal, 2011;

1 We acknowledge that survey response is even more complex than outlined here. A brief review on the psychology of survey response can be found in Schwarz (2008).

Lee, 1993; Lensvelt-Mulders, 2008; Tourangeau & Yan, 2007). Such misreporting to sensitive questions has serious consequences. The prevalence estimates of the sensitive topics are systematically biased and valid analyses on relationships between independent variables and the sensitive behavior cannot be conducted (Bernstein, Chadha, & Montjoy, 2001; Ganster, Hennessey, & Luthans, 1983). Researchers have developed several specific techniques to reduce misreporting on sensitive questions such as wording, framing or randomized response, but empirical research shows that the success of these measures is limited (see Preisendörfer & Wolter, 2014).

We suspect that respondents will be more concerned about their privacy when sensitive topics are involved. Tourangeau and Yan (2007, p. 859) define sensitive questions as "questions that trigger social desirability concerns [and] […] those that are seen as intrusive by the respondents or that raise concerns about the possible repercussions of disclosing the information". Following this definition, our questionnaire encompasses questions with various degrees of sensitivity ranging from low-sensitivity questions on socio-demographics to very sensitive items on group-focused enmity. These items directly point at hostility toward certain groups such as disabled persons, homosexuals, immigrants, Muslims, Jews, homeless, and long-term unemployed. Conforming to such hostile items is clearly socially undesirable, given Western norms. Respondents may experience them as intrusive and fear for their reputation should their attitudes be revealed. More generally, items on political and ethical subjects are considered sensitive (Lensvelt-Mulders, 2008; Stocké, 2007).

As pointed out above, respondents can react to sensitive questions in different ways. First, they can answer truthfully, even if they are aware of their attitude being socially undesirable. Second, they might be reluctant to participate in such a survey (unit nonresponse) or decline answering sensitive questions (item nonresponse). Finally, they might answer sensitive questions but react by adjusting their answers according to what they suspect is socially desirable (misreporting). Survey research has repeatedly shown that sensitive questions increase nonresponse and lead to biased answers.[2] If numeric codes on the questionnaire are perceived as a potential breach in anonymity, all three effects should be enhanced.

The problem of nonresponse can in part be lessened by "a very specific privacy statement" (Lensvelt-Mulders, 2008, p. 467) to attenuate the respondents' concerns for anonymity or confidentiality. On the other hand it is often stated that this privacy statement should not be blatant since "too much emphasize on privacy protection can harm the bond of trust between the respondent and the researcher, resulting in higher nonresponse rates" (Lensvelt-Mulders, 2008, pp. 467f.; see also de Leeuw et al., 2003). However, other research finds only weak effects of different

2 A detailed discussion of reasons for and consequences of respondents' possible reactions to sensitive questions can be found in Tourangeau et al. (2000).

versions of such statements and thus question the extent to which responders actually read confidentiality statements (see, e.g. Tourangeau et al., 2000).

2.3 Previous Empirical Studies

A classic study done by Singer (1978) finds that assuring confidentiality significantly reduces item nonresponse. In their meta-analysis of experimental studies on the effect of confidentiality assurances Singer, von Thurn, and Miller (1995) find a weak but robust positive effect only when sensitive topics were being surveyed. In a large study by Dillman, Singer, Clark, and Treat (1996) there was no difference between different versions of the confidentiality assurance. Ong and Weiss (2000) have shown that assuring anonymity or confidentiality has a strong impact on revealing socially undesirable information. It has also been documented that the sensitivity of the surveyed topic has an impact on the willingness to participate (Couper, Singer, Conrad, & Groves, 2008; Edwards et al., 2002).

Even more closely related to our research, Yang and Yu (2011) examined the effect of personal identifiers on questionnaires. They find that numerical and barcode identifiers on the cover page of a questionnaire both advance nonresponse and reduce socially undesirable answers. They also speculate about sensitive topics being most prone to these adverse effects. On the other hand, there is a number of experiments in surveys on sensitive topics that find no significant effect of a numeric code on unit nonresponse (Campbell & Waters, 1990; Reuband, 1999, 2006, 2015) or on unit nonresponse and reporting behavior (King, 1970; Wildman, 1977). In all these studies, the codes were actual identifiers, i.e. they were unique for every respondent.

2.4 Hypotheses

In conclusion, the review of the literature allows us to formulate these hypotheses:

H1: Numeric codes on questionnaires lead to higher unit nonresponse.

H2: Numeric codes on questionnaires lead to higher item nonresponse in sensitive questions.

H3: Numeric codes on questionnaires lead to answers to sensitive questions biased towards social desirability.

H4: If at least one of the Hypotheses H1-H3 is supported, the results of regressions might be biased.

H5.1-4: An explicit privacy statement in the cover letter addressing the nature of the numeric code might either attenuate or raise unit nonresponse (1), item nonresponse (2), misreporting (3), and biases in regression results (4).

3 Methods

3.1 Survey Design

In February and March 2013, 3,725 paper-and-pencil questionnaires were distributed to randomly selected households in Munich. The sample was generated following a recommendation for drawing local household samples in Bauer (2014). First we randomly selected 712 street sections from a list of 77,218 street sections in Munich. Each had the same probability to be selected. 12,130 households in the selected street sections were manually counted and listed. Based on this list a sample was drawn.

An envelope with the cover letter, the questionnaire, and a stamped return envelope was deposited in the mailboxes of all selected households.[3] The envelope as well as the cover letter and the questionnaire contained a letterhead from the university and were distributed without respondent names, so all respondents were in fact anonymous. The cover letter explained the topic of the survey as "better understanding of political and social developments", asked for the participation of the household member over 18 who had their birthday most recently and emphasized the confidentiality and anonymity of the responses. Furthermore, small bag of jelly babies was added as a little incentive.[4] After two weeks, all households received a reminder postcard. Our methodical proceeding is guided by the tailored design method (Dillman, 2007). In total we received 1,138 questionnaires, which results in a response rate of 30.6%.

As the topic of the study is the analysis of group-focused enmity (GFE) in Munich, the questionnaire covered demographics, housing and neighborhood conditions, and social trends as well as societal and political opinions with an emphasis on GFE. Like in other studies on GFE (Heitmeyer, 2002a, 2002b; Zick et al., 2008) many questions were sensitive since they focus on topics of socially unde-

3 The questionnaire (in German) can be found in the online appendix http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/.

4 Studies suggest that, while it is more effective to use money as an incentive, small presents also have a positive effect (Church, 1993; Edwards et al., 2002; Fick & Diehl, 2013). Given financial restrictions, we decided to give sweets.

sired prejudices and discrimination.[5] The sensitivity of the questions on GFE was particularly suitable for this experiment, since concerns about potentially jeopardized anonymity are expected to influence answering behavior if true answers are discomforting.

3.2 Survey Experiment

To examine the effects of numeric codes on response behavior, we divided the sample into three groups (Table 1). Half of the questionnaires had a code written on the bottom of each page, the other half did not.[6] Half of those questionnaires with code received a cover letter, which explained the numeric codes to respondents: "The digits on the bottom of the questionnaire only facilitate electronic processing and are identical on each questionnaire of this project. They do not permit personal identification" (translated from German).[7] All households within the same street section received one type of questionnaire.

3.3 Data Analysis

All three treatment groups were analyzed regarding unit nonresponse, item nonresponse, and answers to sensitive questions.[8] Since researchers are typically interested in regression results, the groups were also compared in regressions on GFE by interacting all independent variables with the group dummies.

We focused on questions about GFE in the topics of attitudes towards people with disabilities, long-term unemployed persons, homeless people, homosexuals, Muslim and other immigrants, and attitudes towards cultural heterogeneity, anti-Semitism, and National Socialism. All questions from these topics ask respondent to express an attitude towards certain groups. The items range from 1 to 5 and are standardized to ensure that comparisons across items are not dependent on the variance within each item. Very sensitive questions have a lower variance, since

5 For the question wording see Appendix A. Details on data collection and results concerning GFE in Munich can be found in Steinbeißer, Bader, Ganser, & Schmitt (2013). To test the sensitivity of GFE-items, 80 students from the University of Hanover (not the students from Munich who had helped preparing the project) were asked to rate the sensitivity of GFE and some other questions. According to our expectations, GFE-items were rated to be much more sensitive.

6 In previous studies (Campbell & Waters, 1990; King, 1970; Reuband, 1999; 2006; 2015; Wildman, 1977; Yang & Yu, 2011) the codes were located only on the cover page of the questionnaire.

7 The different versions of the cover letter in German can be found in the online appendix http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/.

8 The data are available for scientific purposes from the authors. The R-code can be found at http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/.

Table 1 Variants of the questionnaire in the survey experiment

Treatment Group	Number of Questionnaires	Proportion
Without Numeric Code	1,863	50%
With Numeric Code, Without Notice	931	25%
With Numeric Code and Notice	931	25%
Total	3,725	100%

the truthful answers of only a small fraction of respondents result in the choice of a socially undesirable category. Even if these persons react stronger to a presumed violation of anonymity by the numeric code, smaller effects would be observed, as only a small percentage of respondents is affected. Standardizing the GFE-items allows to analyze variables relative to their own variance.

To test if respondents' reaction to numeric codes depends on the sensitivity of the item, we use a set of less sensitive questions for comparison. The non-sensitive questions cover life-satisfaction and finance, trust in institutions, and good neighborhood. As these questions deal with an individual's personal situation without aiming at morally relevant sentiments towards groups of human beings, we are confident that the urge to give socially acceptable answers is much weaker in these topics than in GFE.

Following Angrist and Pischke (2009) we applied multivariate methods, although we had conducted an experiment, for two reasons: First, there are significant differences between the treatment groups in age, employment status and city district of residence. We consider this an indicator of the randomization procedure, clustered by street sections, not producing a perfectly randomized split. This is why we control for regional and socio-demographic characteristics. Second, the variance in GFE is very large compared to the effect of numeric codes. This masks the effects of the printed codes. Therefore it is necessary to shrink the unexplained variance in GFE by conditioning on suitable explaining variables.

Table 2 gives an overview over mean and standard deviation for all GFE-scales and the used demographic variables. All GFE-items, the non-sensitive items, and their means can be found in the Appendix.

Table 2 Mean for GFE-scales and used demographic variables[a]

Unstandardized GFE-Scales [0,1]	Mean (Standard Deviation)
Xenophobia	0.245 (0.178)
Islamophobia	0.491 (0.250)
Anti-Semitism	0.244 (0.235)
Attitudes Towards Unemployed	0.596 (0.085)
Attitudes Towards Homosexuals	0.225 (0.261)
Attitudes Towards National Socialism	0.142 (0.126)

	Demographic Variables	Mean (Standard Deviation)
	Female	0.536
	German	0.933
	Age	49.3 (17.3)
	Monthly Net Income per Capita (in 1000)	1.722 (1.147)
Religion	Catholic	0.405
	Lutheran-protestant	0.196
	Other	0.031
	None	0.367
Education	No/Junior High School	0.128
	Middle School	0.188
	Adv. Tech. College Qualif.	0.053
	University Qualification	0.130
	University Degree	0.500
Employment	Full-time	0.514
	Regular Part-time	0.119
	Marginal Part-time	0.083
	None	0.283
	Ever Registered as Unemployed	0.337

[a] For metric variables standard deviation in parentheses.

4 Results

4.1 Unit Nonresponse

The response rates in the three treatment groups are very similar and the differences are clearly not significant (see Figure 1 and Table 3). The 3-sample test for equality of proportions (Wilson, 1927) gives a χ^2-value of 0.314 and a p-value of 0.855.

It is save to conclude that the codes with or without notice have a minimal or no effect on unit nonresponse.

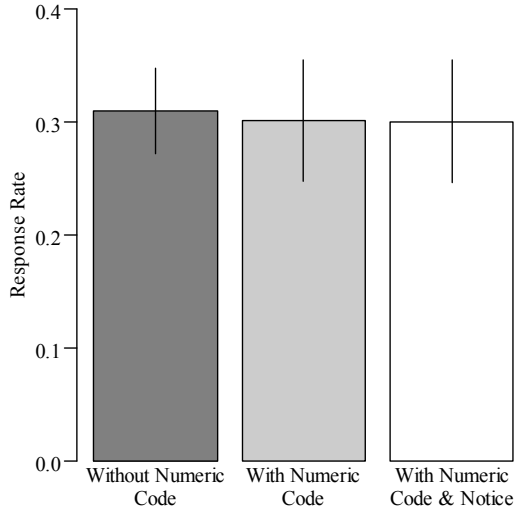

Figure 1 Response rate (percent and 95%-confidence interval) by treatment groups

Table 3 Response by treatment groups

Treatment Group	Number of responses	Response rate
Without Numeric Code	577	31.0%
With Numeric Code, Without Notice	281	30.2%
With Numeric Code and Notice	280	30.1%
Total	1,138	30.6%

4.2 Item Nonresponse

Overall, given the sensitivity of the survey, item nonresponse is pretty low. On average, a sensitive question on GFE was not answered by 13.6 of the 1,138 respondents (1.19%).

Each dot in Figure 2 represents the nonresponse to an item from the questionnaire. For example, the dots marked by arrows represent the item nonresponse to the question "To what extent do you agree with the following statement: The customs and habits of Islam feel creepy to me" (translated from German). In the group without numeric code, the nonresponse rate of this item is 1.04%, while it is higher for respondents who received a questionnaire with numeric code (1.78%) or with numeric code and notice in the cover letter (1.79%).

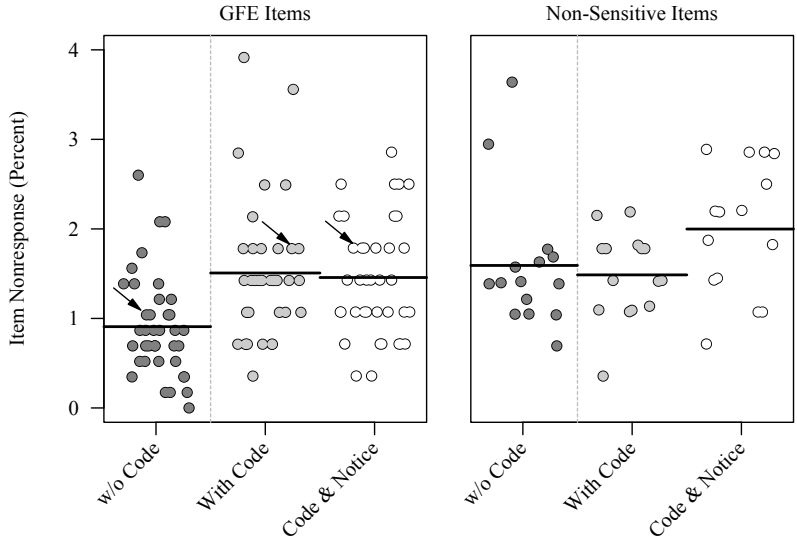

Figure 2 Item nonresponse rates (percent) for GFE-items and non-sensitive items. Black lines represent the average percentage of item nonresponse in each group. Arrows point to the example item in section 4.2 (skepticism about customs and habits of Islam).

Table 4 Average item nonresponse rates for GFE-items and non-sensitive items (p-values in parantheses)

Treatment Group	GFE			Non-Sensitive Items		
Without Numeric Code	0.91%			1.59%		
With Numeric Code, Without Notice	1.51%	(0.059)[a]	(0.032)[b]	1.49%	(0.570)[a]	(0.394)[b]
With Numeric Code and Notice	1.46%	(0.077)[a]		2.00%	(0.268)[a]	

[a] P-values for testing against the group without numeric code.
[b] P-value for testing both groups with numeric code together against the group without numeric code.

The group without numeric code has an average GFE-item nonresponse rate of 0.91%, compared to 1.51% in the group with numeric code and 1.46% in the group with notice in the cover letter (Table 4). The item nonresponse is low in all three groups and the differences are small in absolute numbers, but relatively we see a 66.1% and 60.5% rise in item nonresponse as an effect of the numeric code.

In Figure 2 and Table 4 we compare these results on GFE-items to group differences in nonresponse to non-sensitive items. Surprisingly, the non-sensitive items in general have a higher item nonresponse than the GFE-items. However, the items

show no clear rise in nonresponse due to the numeric code. The group differences in nonresponse to GFE-items are close to significance ($p_{\text{without code vs. code without notice}} = 0.059$, $p_{\text{without code vs. code with notice}} = 0.077$), whereas in case of the non-sensitive items they are clearly nonsignificant ($p_{\text{without code vs. code without notice}} = 0.570$, $p_{\text{without code vs. code with notice}} = 0.268$). If we pool both groups with numeric code and test the difference in item nonresponse between this pooled group and the group without numeric code, the difference turns out to be significant in GFE-items ($p_{\text{without code vs. with code}} = 0.032$) while it is still insignificant in the non-sensitive items ($p_{\text{without code vs. with code}} = 0.394$).[9] A numeric code on the questionnaire does indeed increase item nonresponse but only in sensitive questions. The notice in the cover letter does not influence this effect.

4.3 Misreporting

To analyze the impact of numeric codes on given responses, we estimate group effects for all 38 standardized GFE-items applying a multivariate regression. In order to reduce the residual error all models include sex, age, age squared, religious affiliation, German citizenship, educational status, employment status, income per capita, and city district of residence. This regression on the standardized GFE-items results in parameters for respondents with numeric code but without notice and respondents who received a notice. The group without numeric code serves as reference.

Figure 3 shows the coefficients for the groups from the multivariate regression. Each dot represents the estimated difference in agreement with an item between the groups with numeric codes and the group without numeric code, which serves as reference. As items are standardized, coefficients represent the differences measured in standard deviations of the item. Using the example in section 4.2 regarding skepticism about customs and habits of Islam, respondents who received a questionnaire with a numeric code reported a 0.158 standard deviations lower agreement with this statement. Respondents, who were informed about the use of the numeric code, reported a 0.157 standard deviations lower agreement (see Arrows in Figure 3).

9 To take into account that the answers within respondents are not independent, the test calculates the p-value by applying a Monte Carlo simulation. Values are drawn from a multivariate normal distribution using the variance covariance matrix from a multivariate regression. The multivariate regression contains only the experimental groups as explanatory variables and the nonresponse to each item as dependent variables. As the explanatory variables are all binary, common standard errors can be used. Sociodemographics do not contribute to the explanation of item nonresponse and therefore are not included in the model. As we expect that the numeric code increases nonresponse, we use one-sided tests.

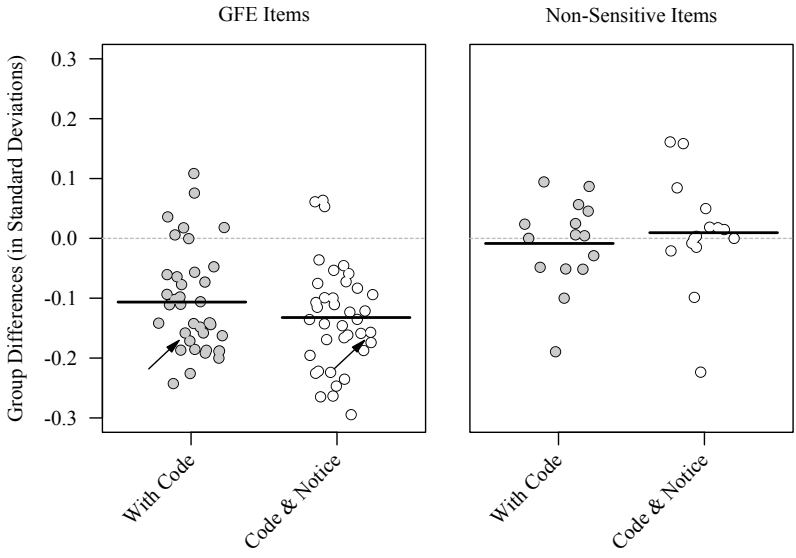

Figure 3 Regression coefficients for standardized GFE-items and standardized non-sensitive items (reference group: without code). Black lines represent the average group differences. Arrows point to the example item in section 4.2 (skepticism about customs and habits of Islam).

Table 5 Average regression coefficients for standardized GFE-items and non-sensitive items (p-values in parantheses)[a]

Treatment Group	GFE		Non-Sensitive Items	
With Numeric Code	-0.106	(0.022)	-0.009	(0.474)
With Numeric Code and Notice	-0.132	(0.011)	0.009	(0.612)

[a] Differences in standard deviations. P-values for testing against the group without numeric code.

All estimates of the multivariate GFE-regression for the groups with numeric code are between -0.294 and 0.108. The majority of coefficients are negative indicating a decrease in socially undesired responses resulting from the scanner code. Respondents with numeric code show lower GFE, i.e. their answers to sensitive questions tend more strongly towards desirability. The average difference over all items on GFE between the group without and the group with numeric code and without notice is -0.106 standard deviations and -0.132 standard deviations to the group with numeric code and notice (Table 5). In case of the non-sensitive items the differences are minimal (-0,009 and 0,009, Figure 3, right panel).

To compare the average effect of numeric codes, significance tests need to account for the dependencies between GFE-items within respondents. The test procedure is based on the variance-covariance matrix of the multivariate regression (similar to the one in the analysis of item nonresponse in 4.2, see footnote 9).[10] On average, the respondents with numeric code show significantly lower GFE ($p_{\text{without code vs. code without notice}} = 0.022$, $p_{\text{without code vs. code with notice}} = 0.011$). Respondents with numeric codes gave more socially desirable answers to sensitive questions. In case of the non-sensitive items the differences are insignificant ($p_{\text{without code vs. code without notice}} = 0.474$, $p_{\text{without code vs. code with notice}} = 0.612$).

We can conclude that the numeric code discourages respondents to give socially undesirable answers to sensitive questions but seems to have no impact concerning non-sensitive questions. While the effect of an explanation of the code's usage is not definite, it is safe to state that it has no positive effect.

4.4 Impact on GFE-regression results

So far, the results show higher item nonresponse and more socially desired answers in the treatment groups with numeric code. As most studies are interested in relationships between variables, we calculate a typical GFE-model for xenophobia and compare how the numeric code and the notice affect the regression coefficients.[11] All variables are interacted with the group dummies "code without notice" and "code with notice". If numeric codes do not influence respondents' behavior, the interactions should be insignificant. If numeric codes influence regression results, estimated interaction parameters should show significant effects.

The interaction coefficients and confidence intervals of the model are shown in Figure 4. With a critical value of 0.05, 26 explaining variables, and two interaction groups one would expect 2.6 parameters to be significant by chance. Three interactions terms with the numeric code group without notice became significant (gender, catholic religion and other religion). Interactions with notice did not exceed the critical value. All in all there are not more significant parameters than expected, which indicates that code and notice do not affect the regression results systematically.

10 To test the standardized non-sensitive items we use the absolute coefficients, as the pooling of the items would otherwise influence the result. However, pooling is not possible, as there is no general expectation what kind of answer is more socially undesirable. For the standardized GFE-items the normal coefficients are used. P-values are calculated by drawing values from a multivariate normal distribution with the covariance matrix taken from the multivariate regressions for standardized GFE-items and standardized non-sensitive items. The values (GFE) or absolute values (non-sensitive items) from the drawn sample serve as error distribution. The multivariate normal distribution accounts for the dependencies between items.

11 The variables used in this regression are guided by models in the report on the survey Steinbeißer et al. (2013).

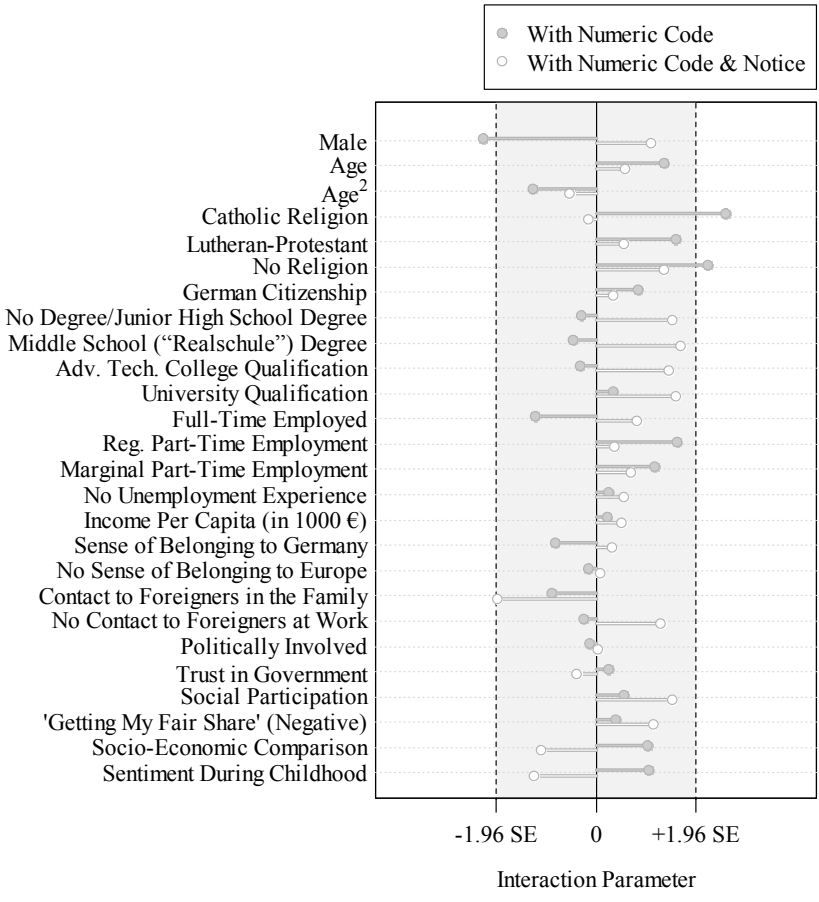

Figure 4 Interaction parameters of the treatment group dummy variables (ref: no numeric code) with several typical independent variables and 95%-confidence interval from a regression for xenophobia (n = 706).

Several other regression models support this result. We estimate additional models with the same independent variables for islamophobia, anti-Semitism and attitudes towards unemployed, homosexuality and National Socialism. When looking at the distribution of the interaction parameters, there are never more than four significant interaction parameters per model. This is consistent with the distribution of t-values for all interactions. If there is no relationship and the interaction parameters descend from a random distribution, t-values should converge to a standard normal distribution. The interaction terms are close to a standard normal distribution. The Kolmogorov-Smirnov test, Shapiro-Wilk test and Anderson-Darling test show no significant deviation for the group without notice ($p_{KS} = 0.676$, $p_{SW} = 0.614$,

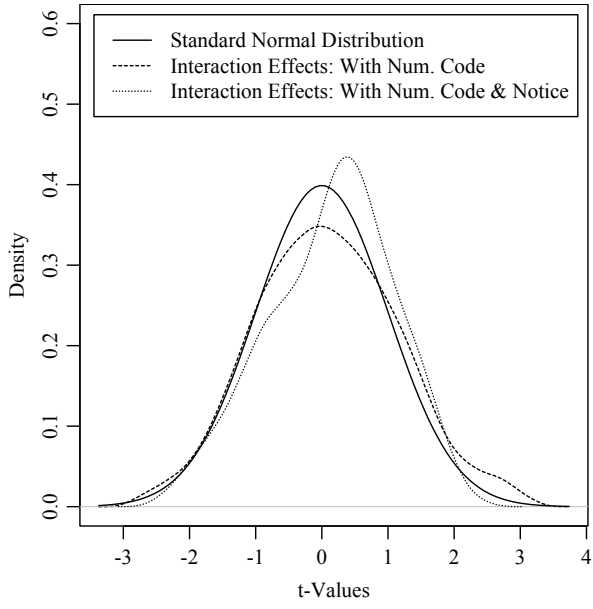

Figure 5 Distribution of 156 interaction parameters for treatment group dummy variables (ref: no numeric code) with independent variables from six regression models for GFE-variables (kernel density estimation).

$p_{AD} = 0.556$). In the group with notice we find a significant deviation with the Kolmogorov-Smirnov test ($p_{KS} = 0.007$) at the 1% significance level and with the Shapiro-Wilk test and Anderson-Darling test, at the 10% level ($p_{SW} = 0.092$, $p_{AD} = 0.052$). This can also be seen in Figure 5. While the peak of the t-value distribution for the group without notice is very close to zero, the peak for the group with notice is not. The test supports the visual impression, that the distribution of the group with notice is not normal. Given these results, an effect on the regression result, caused by the combination using a numeric code and pointing respondents to the code, can be suspected.

5 Conclusion and implications

We hypothesized that numeric codes on a questionnaire might influence respondents' behavior. Such a code could induce privacy concerns in the respondents. Respondents can possibly react by refusing to answer the complete questionnaire (H1), by skipping sensitive questions (H2) or by giving biased answers to sensitive

questions (H3). If the code did have such an impact on respondents' behavior, this might also bias typical regression results (H4). These effects on unit nonresponse, item nonresponse, misreporting, and consequently on regression results can be attenuated or raised by a statement on the code's usage in the cover letter (H5.1, H5.2, H5.3, H5.4). To test these hypotheses a survey experiment was conducted: We printed a numeric code on half of the questionnaires and half of the respondents with a numeric code were informed about it in the cover letter.

Although the response rate was slightly lower in the groups with numeric code, the differences were far from significant. In line with other studies (Campbell & Waters, 1990; King, 1970; Reuband, 1999, 2006, 2015; Wildman, 1977), H1 and H5.1 are not supported.

Respondents with a numeric code had a significantly higher item nonresponse rate in sensitive questions on group-focused enmity, but no higher item nonresponse in non-sensitive items. Explaining the numeric values on the questionnaire in the cover letter did not have an influence on these relationships. H2 is supported, but there is no support for H5.2.

A significant misreporting bias towards socially desired answers was found in both treatment groups. An explanation of the numeric code appears to have no positive effect, however, results give no clear indication whether the notice introduces bias or not. H3 can be confirmed, but there is no support for H5.3.

The influence of these differences on regression results was insignificant. H4 cannot be confirmed. However, if the numeric code is addressed in the cover letter, there is some indication that regression coefficients might be affected, as the bias in single variables seems to accumulate. While the result is ambiguous, there is some support for H5.4.

In contrast to other studies (King, 1970; Wildman, 1977), we find at least some systematic differences in reporting behavior regarding item nonresponse and misreporting. The zero results of older studies might be due to limitations in the sample size and in the statistical techniques. Our survey uses a bigger sample, we apply tests based on multiple variables and we are able to reduce the unexplained variance by adjusting for suitable sociodemographic variables. The results are in accordance with Yang and Yu (2011): respondents who had a questionnaire with a scanner code gave more socially desirable answers. Given that non-sensitive items were not affected, it seems reasonable to assume that respondents reacted to a perceived breach in anonymity. We suspect that there are two ways in which a smaller item mean emerges. Several respondents reacted to the numeric code and adjusted their response a little in direction of a social desirable answer. In addition, a small group, who would honestly give extreme answers, completely changed their response to the opposite socially desirable extreme. We think that both influences are possible; however, our data does not allow distinguishing between these effects.

Given that optical mark recognition software very often uses numeric codes, we also want to give recommendations for survey methodologists. First, we recommend to avoid an explicit statement on that code in the cover letter (if legally and ethically justifiable) as it does not ease biases. Given our results, a numeric code can be problematic if there is a specific interest in descriptive results on a sensitive topic (like unemployment experiences). It is also possible that some groups of respondents react strongly to a numeric code on the questionnaire. On the other hand, the effects are small compared to other sources of bias like selective nonresponse to surveys (Schnell, 1997) and mismatch of answers and behavior (see e.g. Diekmann & Preisendörfer, 1992).

Second, researchers need to trade off the potential biases brought about by a numeric code and the importance of the code on the questionnaire for the research project. In case of automated capture of questionnaires instead of manual data entry, numeric codes will save money that can be used to reduce selective unit nonresponse.

While this study focuses on scanner codes, there are other potential purposes of such codes. Numbers on questionnaires can be utilized as panel identifiers, treatment group identifiers in survey experiments, and for region mapping. As far as the conditions for such codes are similar, our results should largely be transferable. If numeric codes are applied thoughtfully, their benefit for the usage of statistical procedures to reduce errors could well outweigh the code's negative effect.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton Univ. Press.

Babbie, E. (2013). *The Practice of Social Research: Thirteenth Edition*. Belmont: Wadsworth.

Barnett, J. (1998). Sensitive Questions and Response Effects: An Evaluation. *Journal of Managerial Psychology, 13*(1/2), 63-76.

Barton, A. H. (1958). Asking the Embarrassing Question. *Public Opinion Quarterly, 22*(1), 67-68.

Bauer, J. J. (2014). *Verzerrungen in Random-Route Stichproben und Lösungsansätze: Vortrag ETH Zürich*. Retrieved from http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/index.html.

Benson, L. E. (1941). Studies in Secret-Ballot Technique. *Public Opinion Quarterly, 5,* 79-82.

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting Voting. Why it Happens and Why it Matters. *Public Opinion Quarterly, 65*(1), 22-44.

Beyer, H., & Krumpal, I. (2010). „Aber es gibt keine Antisemiten mehr": Eine experimentelle Studie zur Kommunikationslatenz antisemitischer Einstellungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62*(4), 681-705.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions*. San Francisco: Wiley.

Campbell, M. J., & Waters, W. E. (1990). Does Anonymity Increase Response Rate in Postal Questionnaire Surveys about Sensitive Subjects? A Randomised Trial. *Journal of Epidemiology and Community Health, 44*(1), 75-76.

Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly, 57*(1), 62-79.

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation. *Journal of Official Statistics, 24*(2), 255-275.

de Leeuw, E. D., & Hox, J. (2008). Self-administred Questionnaires: Mail Surveys and Other Applications. In E. D. de Leeuw, J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 239-263). New York, London: Lawrence Erlbaum Associates.

de Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics, 19*(2), 153-176.

de Rada, V. D. (2005). Influence of Questionnaire Design on Response to Mail Surveys. *International Journal of Social Research Methodology, 8*(1), 61-78. doi:10.1080/1364557021000025991

Diekmann, A., & Preisendörfer, P. (1992). Persönliches Umweltverhalten: Diskrepanzen zwischen Anspruch und Wirklichkeit. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 44*(2), 226-251.

Dillman, D. A. (1991). The Design and Administration of Mail Surveys. *Annual Review of Sociology, 17,* 225-249.

Dillman, D. A. (2007). *Mail and Internet Surveys: The Taylored Design Method* (2nd Edition). New York: John Wiley & Sons.

Dillman, D. A. (2008). The Logic and Psychology of Constructing Questionnaires. In E. D. de Leeuw, J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 161-175). New York, London: Lawrence Erlbaum Associates.

Dillman, D. A., Singer, E., Clark, J. R., & Treat, J. B. (1996). Effects of Benefits Appeals, Mandatory Appeals, and Variations in Statements ofConfidentiality on Completion Rates for Census Questionnaires. *Public Opinion Quarterly, 60*(3), 376-389.

Edwards, P., Roberts, I., Clarke, M., Di Giusseppi, C., Pratap, S., Wentz, R., & Kwan, I. (2002). Increasing Response Rates to Postal Questionnaires: Systematic Review. *British Medical Journal, 324,* 1183-1192. doi:10.1136/bmj.324.7347.1183

Fick, P., & Diehl, C. (2013). Incentivierungsstrategien bei Minderheitsangehörigen. Ergebnisse eines Methodenexperiments. *methoden, daten, analysen, 7*(1), 59-88.

Ganster, D. C., Hennessey, H. W., & Luthans, F. (1983). Social Desirability Response Effects: Three Alternative Models. *Academy of Management Journal, 26*(2), 221-331.

Heitmeyer, W. (Ed.). (2002a). *Deutsche Zustände: Folge 1*. Frankfurt: Suhrkamp.

Heitmeyer, W. (2002b). Gruppenbezogene Menschenfeindlichkeit. Die theoretische Konzeption und erste empirische Ergebnisse [Group-focused enmity. Theoretical Conception and First Empirical Results]. In W. Heitmeyer (Ed.), *Deutsche Zustände. Folge 1* (pp. 15-36). Frankfurt: Suhrkamp.

Hyman, H. (1944). Do They Tell the Truth? *Public Opinion Quarterly, 8,* 557-559.

King, F. W. (1970). Anonymous versus Identifiable Questionnaires in Drug Usage Surveys. *American Psychologist, 25*(10), 982-985.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys. The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly, 72*(5), 847-865.

Krumpal, I. (2011). Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality & Quantity, 47,* 2025-2047.

Lee, R. M. (1993). *Doing Research on Sensitive Topics*. Thousand Oaks: SAGE.

Lensvelt-Mulders, G. (2008). Surveying Sensitive Topics. In E. D. de Leeuw, J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 461-478). New York, London: Lawrence Erlbaum Associates.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

Ong, A. D., & Weiss, D. J. (2000). The Impact of Anonymity on Responses to Sensitive Questions. *Journal of Applied Social Psychology, 30*(8), 1691-1708.

Preisendörfer, P., & Wolter, F. (2014). Who Is Telling the Truth? A Validation Study on Determinants of Response Behavior in Surveys. *Public Opinion Quarterly, 78*(1), 126-146.

Reuband, K.-H. (1999). Anonyme und nicht-anonyme postalische Bevölkerungbefragungen. *Planung & Analyse, 26*(1), 56-58.

Reuband, K.-H. (2006). Postalische Befragungen alter Menschen: Kooperationsverhalten, Beantwortungsstrategien und Qualität der Antworten. *ZA-Information / Zentralarchiv für Empirische Sozialforschung, 59,* 100-127.

Reuband, K.-H. (2015). Ausschöpfung und Nonresponse Bias in postalischen Befragungen. In J. Schupp & C. Wolf (Eds.), *Schriftenreihe der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute. Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen* (pp. 209-251). Wiesbaden: Springer VS.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A Meta-Analytic Study of Social Desirability Distortion in Computer-Administered Questionnaires, Traditional Questionnaiers, and Interviews. *Journal of Applied Psychology, 84*(5), 754-775.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmass, Entwicklung und Ursachen*. Opladen: Leske + Budrich.

Schwarz, N. (2008). The Psychology of Survey Response. In W. Donsbach & M. W. Traugott (Eds.), *The SAGE Handbook of Public Opinion Research* (pp. 374-387). Los Angeles, London: SAGE.

Singer, E. (1978). Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys. *American Sociological Review, 43*(2), 144-162.

Singer, E., von Thurn, D. R., & Miller, E. R. (1995). Confidentiality Assurances and Response: A Quantitative Review of the Experimental Literature. *Public Opinion Quarterly, 59*(1), 66. doi:10.1086/269458

Steinbeißer, D., Bader, F., Ganser, C., & Schmitt, L. (2013). *Gruppenbezogene Menschenfeindlichkeit in München: Forschungsbericht des Instituts für Soziologie der Ludwig-Maximilians-Universität München*. Retrieved from http://www.ls4.soziologie.uni-muenchen.de/forschung/aeltere_projekte/gmf/bericht_gmf_18_10_2013.pdf

Stocké, V. (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Rational-Choice Theorie und des Modells der Frame-Selektion. *Zeitschrift für Soziologie, 33*(4), 303-320.

Stocké, V. (2007). The Interdependence of Determinants for the Strength and Direction of Social Desirability Bias in Racial Attitude Surveys. *Journal of Official Statistics, 23*(4), 493-514.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge university press.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin, 133*(5), 859-883.

Warwick, D. P., & Lininger, C. A. (1975). *The Sample Survey: Theory and Practice*. New York: McGraw-Hill.

Wildman, R. C. (1977). Effects of Anonymity and Social Setting on Survey Responses. *Public Opinion Quarterly, 41*(4), 74-79.

Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association, 22*(158), 209-212. doi:10.1080/01621459.1927.10502953

Wolter, F. (2012). *Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik*. Wiesbaden: Springer-VS.

Yang, M.-L., & Yu, R.-R. (2011). Effects of Identifiers in Mail Surveys. *Field Methods, 23*(3), 243-265. doi:10.1177/1525822X11399401

Zick, A., Wolf, C., Küpper, B., Davidov, E., Schmidt, P., & Heitmeyer, W. (2008). The Syndrome of Group-Focused Enmity: The Interrelation of Prejudices Tested with Multiple Cross-Sectional and Panel Data. *Journal of Social Issues, 64*(2), 363-383. doi:10.1111/j.1540-4560.2008.00566.x

Appendix

A Items on GFE

In the items on GFE the survey respondents were asked to gradually agree or disagree with these statements on a scale of 1 to 5 ("Totally disagree", "Rather disagree", "Neither agree nor disagree", "Rather agree", "Totally agree"). The items we used are:[12]

GFE-Items [1-low approval, 5-high approval]	Mean
To what extent do you agree with the following statements concerning disabled persons?	
In Germany, more should be done for disabled persons.[a]	3.845
I find many demands by disabled persons excessive.	2.120
Disabled persons receive too many privileges.	1.695
The following is about opinions on unemployed persons. In your opinion, to what extent do the following statements apply?	
Most unemployed persons make an effort to find a job.[a]	3.006
Unemployed persons who don't find a job after longer search are themselves to blame.	2.459
I find it outrageous how permanently unemployed persons live a comfortable life at the society's expense.	2.635
How some people systematically dodge work makes me angry.	3.857
Permanently unemployed persons should receive more support so they can find back to a working life.[a]	3.614
To what extent do you agree with the following statements concerning homeless persons?	
Most homeless persons have gotten into this situation through no fault of their own.[a]	3.178
Begging homeless persons should be removed from pedestrian areas.	2.666
Most homeless persons are disinclined to work.	2.452
To what extent do you agree with the following statements concerning homosexuality?	
Homosexuality is immoral.	1.453
Marriages between two women or, two men, respectively, should be allowed.[a]	3.994
Adopting children should stay forbidden to same-gender couples.	2.263

[a] Item scale is reversed in all calculations.

12 The complete questionnaire (in German) can be found in the online appendix http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/.

GFE-Items [1-low approval, 5-high approval]	Mean
Now we would like to know to what extent you agree with the following statements.	
Jewish culture is an important part of Germany.[a]	3.610
What our country needs today is a tough and forceful assertion of German interests towards other countries.	2.581
A National Socialist dictatorship must never be allowed to happen again.[a, b]	4.911
National Socialism also had its good sides.	1.481
Like in nature, the strongest should always prevail in society.	1.683
Actually, Germans are, by nature, superior compared to other peoples.	1.335
Even today the influence of Jews is too great.	2.014
We should have a leader who governs Germany with a strong hand for the good of all.	1.272
Jews simply have something special and peculiar about them and don't quite fit with us.	1.526
In the past months there has been a lot of debate in the public about immigration and integration. Therefore we are interested in the extent to which you agree with the following statements.	
Muslim culture fits with Germany well.[a]	2.549
Foreigners only come here to exploit our welfare state.	2.529
Naturalization of immigrated foreigners should be facilitated.[a]	3.010
The building of mosques enriches cultural life in Munich.[a]	2.888
When jobs become scarce, foreigners should be sent back home.	2.008
There are too many foreigners in our neighborhood.	2.025
An employer should be allowed to hire only Germans.	1.608
Having a variety of different religions is good for a country.[a]	3.804
I would only reluctantly register my child in a kindergarden/a school with many foreign children.	2.749
The customs and habits of Islam feel creepy to me.	2.847
Foreigners should leave Germany as fast as possible.	1.422
Foreigners living here threaten my own financial situation.	1.365
Munich is superalienated by foreigners to a dangerous degree.	1.779
In our society, too little regard is taken for minorities.	2.884
We have to protect our culture against the influence of other cultures.	2.499
There are too many Muslims in Germany.	2.425

[a] Item scale is reversed in all calculations.

[b] Due to the extremely high approval rate, there was almost no variation within the item. It is therefore not used for analyses which focus on single items (sections 4.2 and 4.3). It is, however, part of the National Socialism scale (Table 2 and section 4.4).

B Non-Sensitive Items

These items are used as test group in order to notice biased answers to less sensitive items.

Non-Sensitive Items	Mean
Please state how comfortable you are in your neighborhood.	
[very uncomfortable-1, very comfortable-5]	4.188
In your personal opinion, to what extent do the following statements apply to your nearer living environment?	
[low approval-1, high approval-5]	
The people here help each other.	3.380
The people here know each other well.	2.997
The people here get along well with each other.	3.719
All things considered, how satisfied are you with your life as a whole these days?	
[very unsatisfied-1, very satisfied-5]	3.975
How do you rate your current financial situation?	
[very bad-1, very good-5]	3.628
How much of what you want can you afford?	
[nearly nothing/nothing-1, nearly everything/everything-5]	3.448
Are you worried about your job?	
[very much-1, not at all-5, I don't work (anymore)-Missing]	4.606
How much trust do you have in…	
[very little-1, very much-5, I don't know-Missing]	
… the Bundestag?	2.919
… the German economy?	3.444
… churches?	2.283
… courts/the legal system?	3.329
… schools/the educational system?	3.110
… the current federal government?	2.860
… the police?	3.492